**PROJECT STATUS UPDATE**

The primary task at first is to prepare the data. The data provided by the IMDB website is in text format (.list) scattered over 49 files. The common factor amongst all the files is the name of the movie. The initial work being done is to identify the relevant features out of the forty nine files available, and form one data frame on the basis of the movie name, consisting of the features and the target variable.

Some of the features such as alternate-versions, biographies, crazy-credits, goofs are being dropped on the basis that they are in the form of full text and will require different sort of natural language processing techniques which is beyond the scope of this project.

Some features such as name, title, costume designers etc are being dropped as features on the basis that it is obvious that they have no contribution towards the rating of a movie.

The features that are left after implementing the above considerations will be used.

Q 1: The baseline model used will be a classifier (on the basis of performance a particular model will ne chosen). The target variable (i.e. the movie rating) will be split into two categories: greater than 5 and lesser than 5.The initial model will predict whether the rating for a movie will be greater than 5 or below 5 based on its features.

Q2: The baseline model will be improved in the following stages:

1. The first improvement will increase the number of values for the target variable to four categories:

Movie rating<2.5, 2.5<Movie rating rating<5, 5<Movie Rating<7.5, 7.5<Movie Rating

1. If the results for the previous models are satisfying, another improvement will be to further increase the number of possible values for the target variable to ten possible values thus bringing the predicted value closer to the true value for the target variable.

Movie Rating<1

2<Movie Rating<3

3<Movie Rating<4

4<Movie Rating<5

5<Movie Rating<6

6<Movie Rating<7

7<Movie Rating<8

8<Movie Rating<9

9<Movie Rating

Q3: The models implemented will be compared on the basis of metrics used for classification such as Accuracy and other relevant metrics.